The UMLS Knowledge Source Server: A Versatile Internet-Based Research Tool

Alexa T. McCray, Amir M. Razi, Anantha K. Bangalore, Allen C. Browne, P. Zoe Stavri
National Library of Medicine
Bethesda, MD 20894

The National Library of Medicine's Unified Medical Language System (UMLS) project regularly distributes a set of Knowledge Sources to the research community. In 1995 the UMLS data were made available for the first time through the Internet-based UMLS Knowledge Source Server. The server can be accessed through three different client interfaces. The World Wide Web interface allows users to browse and explore the data and to see how those data are organized in the UMLS. The command-line interface is best suited for batch processing, and the application programming interface allows developers at remote sites to embed calls in their application programs to the Knowledge Source Server.

INTRODUCTION

The National Library of Medicine's Unified Medical Language System (UMLS) project regularly distributes a set of Knowledge Sources to the research community. The first version of the Knowledge Sources was released in the fall of 1990. In January 1996 the seventh experimental edition was released, containing four Knowledge Sources together with ancillary lexical programs for managing linguistic variation in biomedical terminologies.

In 1995 the UMLS data were made available for the first time through the Internet-based UMLS Knowledge Source Server. The server can be accessed through three different client interfaces. The World Wide Web interface allows users to browse and explore the data and to see how those data are organized in the UMLS. Users can request information about a particular concept, for example, retrieving all its synonyms, its definition, its semantic type, and all the other concepts that are saliently related to it in the Metathesaurus. Users can navigate the Semantic Network, exploring its structure, and they can, for example, retrieve all the concepts that refer to medical devices or diseases in the Metathesaurus. The SPECIALIST Lexicon can be searched, and syntactic and morphologic information about the lexical item will be displayed, together with a link to a Metathesaurus definition if there is one. The Web interface to the Information Sources Map (ISM) currently allows users to see the ISM description for each of the databases and also to browse a sample record for each database.

The command-line interface to the server is best suited for batch processing. Researchers can submit a list of terms to the server to see if they can be found in the UMLS; they can search for various attributes of the terms that are found; and they can filter the results, limiting the result set by attribute, for example, to just those terms that have a particular semantic type or a particular lexical tag.

The Application Programming Interface (API) allows developers at remote sites to embed calls in their application programs to the Knowledge Source Server, thereby accessing the UMLS data directly over the Internet. The API has been designed to be simple, consisting of functions for establishing connections to the server, posting queries, and retrieving the query results.

RESEARCH INVESTIGATIONS USING THE UMLS KNOWLEDGE SOURCES

Currently, over six hundred individuals or institutions receive the UMLS Knowledge Sources on CD-ROM. A much smaller number of these (approximately eighty individuals or institutions, at this writing), has requested access to the UMLS Knowledge Source Server. Recent research has investigated the use of the UMLS for clinical applications as well as for a variety of information retrieval tasks. The Internet-based Knowledge Source Server is potentially of use to these types of investigations and many others where the research depends on accurate access to UMLS data. If the researcher has adequate access to the Internet, the benefit of using the server could be quite significant, since it means that the individual investigator does not need to invest the time and effort in designing and implementing local programs to manipulate and access the extensive UMLS data distributed on CD-ROM. An additional potential benefit of using the Knowledge Source Server is that it should help ensure comparability of results. Since the Knowledge Source Server has

0195—4210/96/$5.00

only recently become available, it is not surprising that studies such as the following which have depended on access to the UMLS have, in most cases, involved the development of local routines to extract and then match against the UMLS data. The Knowledge Source Server, and in particular, its normalized indexes, which abstract away from a range of linguistic variation and which are intended to optimize searching and matching on UMLS concepts, could be useful in future studies of this kind.

The studies described below are some examples of the type of research investigations that have been conducted using the UMLS Knowledge Sources. The coverage of Metathesaurus terminology for use in the clinical setting was investigated in a study which automatically substituted Metathesaurus concepts for terms in physician diagnostic statements [1]. The Metathesaurus was studied in a comparative analysis of four coding schemes used to capture patient problem lists [2]. Surgical operative reports were parsed and the UMLS Semantic Network was used to disambiguate terminology in narrative text [3], and a urology subset of the Metathesaurus was created by exploiting the inter-concept links represented in the UMLS [4]. The Metathesaurus is being used to index and organize an image database in three clinical domains: oncology, gastroenterology, and clinical pathology [5].

UMLS knowledge is heavily used to capture patient care information at the Columbia-Presbyterian Medical Center, where it forms a part of the Columbia Medical Entities Dictionary [6]. The Metathesaurus and the Semantic Network have been used in a range of information storage and retrieval experiments and systems. Conceptual models for use in information retrieval systems have been created by exploring the concepts, relations, and semantic types represented in the UMLS [7,8,9]. The use of the UMLS to augment the selection of search terms in information retrieval systems has been investigated by many researchers, e.g., [10-12]. The UMLS Knowledge Source Server, both in its beta and final release versions, has been actively used in [5,11,12].

IMPLEMENTATION

The implementation of the UMLS Knowledge Source Server is based on the client-server paradigm. The system architecture consists of three types of client interfaces, a main server, and four subordinate servers, one for each of the Knowledge Sources (see Figure 1).

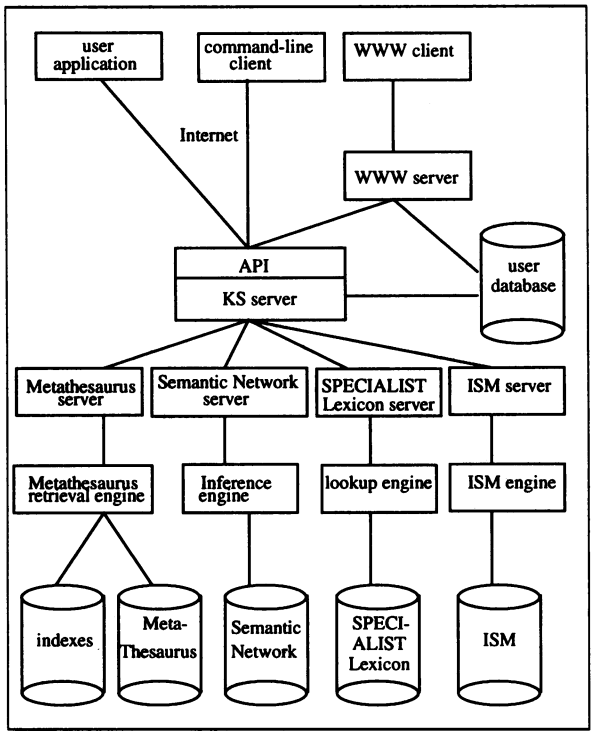

Figure 1. UMLS Knowledge Source Server System Architecture

Server

Each server consists of a master process and several slave processes. The master process receives the client's request and initiates a concurrent process for responding to that request. The concurrent process dies after it retrieves the requested information. If the number of concurrent requests that the server receives grows beyond the limits of our current system, future implementations might involve a transaction oriented processing of the requests based on a fixed number of concurrent processes. Each of the subordinate servers consists of a back-end search engine, the UMLS data files, and multiple index structures. The implementation of the indexes is based on a prefix B+ tree method. Each index tree is a k-branch tree with the non-leaf nodes providing indexes into the range of keys represented by each of the children nodes and their descendants which eventually connect to leaf nodes. The leaf nodes contain pointers to the location of the data identified by the index. The prefix B+ tree generates intermediate prefix keys which allows for prefix searches given a partial key.

The Metathesaurus data tables are indexed on a variety of fields including the concept, term and string

unique identifers, and the string itself. Based on the individual indexes, a master index is then created where, for a given term, direct pointers to all respective data records in relevant Metathesaurus data files are provided. The master index significantly speeds up the retrieval process by eliminating individual index look ups when asking for a range of information about a term, such as its definition, its semantic type, its lexical tags, etc. Indexes are also generated for several field values, allowing efficient retrieval of a list of terms that all share a particular attribute value, such as all those concepts having the same semantic type. The word index files are first indexed by creating a file of words each pointing to the file location of the string containing that word. An index is generated for that file, locating the beginning of the blocks corresponding to all strings containing each word.

Clients

Application programs can be built by embedding online access to the UMLS data through the published API. A simple algorithm for accessing the data would consist of the following steps: a) open connection to the UMLS Knowledge Source Server; b) post a formulated query; c) while there is a response on the connection, get the next line of the response; d) close the connection.

The command-line interface consists of a call to the server with some number of options with their arguments (which themselves consist of options with arguments). For example, a query to the Metathesaurus for all terms with the lexical tag "eponym" would be expressed as follows: "ks -meta -alt eponym". Thus, the argument to "-meta" is "-alt eponym", and the argument to "-alt" is "eponym".

The Web interface is implemented as a series of Common Gateway Interface (CGI) scripts and uses a socket based API to connect to the back-end server. Information is organized in such a way as to be useful to both the novice and the expert. Help on using the various features of the interface is provided on each page. After the user submits a query, this is sent to the back-end through a CGI script. The output is displayed either as fielded ASCII or HTML, or as a scrolling pick list, depending on the type of query made. For example, if the query involves one of the word indexes, then several candidate terms will be returned in a scrolling list. The user can choose an item from the list which is then copied into a type-in window, thereby freeing the user from re-typing the item when further information is desired.

FLEXIBLE ACCESS TO UMLS TERMINOLOGY

Word and Term Indexes

In addition to direct access through the strings and lexical variants provided by the Metathesaurus, the Knowledge Source Server offers access to Metathesaurus terminology through several special indexes. The word indexes allow users to search for terms containing a given word. The normalized word index ignores punctuation, word order and inflectional variation, retrieving items that would otherwise not be found. For example, a search through the simple word index for terms containing the word "bacterium" yields 10 terms, including "gram negative bacterium" and "gram positive bacterium". The same search through the normalized word index yields a total of sixty-seven terms, including "aerobic bacteria", and "gram negative anaerobic bacteria". The word indexes allow for multi-word look up and word truncation. For example, a search through the normalized word index using the words "respiratory" and "function" returns the terms "decreased respiratory function" and "brain stem respiratory center function", among several others. Right truncation is particularly useful when a user is not sure of the spelling of a term or does not wish to type a very long term.

Exploring UMLS data through the Knowledge Source Server

An example best illustrates the type of concept information available in the UMLS that can easily be found by using the Knowledge Source Server. If the user is interested in the concept "attention deficit disorder", a search for basic concept information will reveal that this concept appears in over a dozen thesauri, including the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), SNOMED International, Thesaurus of Psychological Index Terms, the International Classification of Diseases (ICD9-CM), and the Medical Subject Headings (MeSH). The concept has a variety of names (synonyms), including "hyperkinetic syndrome" and "attention deficit disorder with hyperactivity". Since this is a MeSH term, and since the Metathesaurus currently includes French, Spanish, Portuguese, and German MeSH terms, those translations are also listed; e.g., for German, "Konzentrationsstoerung mit Hyperaktivitaet". Three sources have a definition for this concept (e.g., Dorland's Illustrated Medical Dictionary: "...a controversial childhood mental disorder with onset before age seven..."). Dorland's also lists additional alternative names for this sydrome, including "hyperkinetic

reaction" and "hyperactive child syndrome". Finally, the semantic type assigned to the concept is "Mental or Behavioral Dysfunction".

A host of terms related to "attention deficit disorder" can be found. Searching through the normalized word index for terms that are purely lexically related yields a list that includes three variants of "attention-deficit / hyperactivity disorder": one qualified by "combined type", another by "predominantly hypertensive-impulsive type", and another by "predominantly inattentive type". Searching through the related information option reveals that in DSM-IV the parent of the term is "attention-deficit and disruptive behavior disorders"; in SNOMED International its parent is "disruptive behavior disorders"; and in MeSH its parent is "child behavior disorders". Terms related to "attention deficit disorder", but not in a child/parent relationship in a particular thesaurus, include "hyperkinesis", "impulsive behavior", "distractibility", and "short attention span".

By looking at the locator information, the user sees that "attention deficit disorder" is found as "hyperactivity of children" in Online Mendelian Inheritance in Man. The locator information also reports that during the past ten years over 1,600 MEDLINE citation records have included this concept as a major topic of discussion. Further, by choosing the co-occurrence; grouped by semantic type option which displays the terms that co-occur with this main concept in MEDLINE citation records, the user sees that the most frequently co-occurring semantic type is "Organic Chemical". By far the most frequently co-occurring organic chemical listed there is "methylphenidate" (273 instances), with the next most frequent co-occurrence being "central nervous system stimulants" (38 instances). In the Web interface, it is a simple matter to explore the term "methylphenidate" further. Clicking on the term in the co-occurrence window will return the user to the basic concept information for the term. The definition for "methylphenidate" reads "A central nervous system stimulant used most commonly in the treatment of attention-deficit disorders in children..." So, it is clear that what has been codified by this definition is also reflected in the scientific literature on this topic.

It is possible to search for all the other mental disorders that are included in the Metathesaurus, by choosing the option search by attribute; all concepts with a semantic type. Choosing "Mental or Behavioral Dysfunction" yields a list of over 2,000 additional mental disorders. Browsing the Semantic Network reveals that "Mental or Behavioral Dysfunction" can be related to other semantic types in a variety of ways. For example, it "has_manifestation Finding"; it "has_manifestation Behavior"; and it may be "treated_by a Pharmacologic Substance." All of these relationships are reasonable and are, in fact, reflected by the type of related terms that are found in the Metathesaurus, as discussed above. For example "attention deficit disorder" (Mental or Behavioral Function) is manifested by "short attention span" (Finding) and "impulsive behavior" (Individual Behavior), and it is treated by "methylphenidate" (Pharmacologic Substance).

In addition to focusing on a single term, as has been done in the above example, users can explore the UMLS in batch mode. Thus, it is possible, through the command-line option, to submit an entire file of terms and query the Knowledge Source server for one or more attributes of those terms. For example, "ks -meta -f myterms -c -def -cst -anc -des" searches the Metathesaurus (-meta) for information about the terms in a file (-f myterms), and for each of the terms, it finds concept information (-c), the definition (-def), the semantic types (-cst), and all ancestors (-anc) and descendants (-des).

These examples illustrate some of the information that is readily available by browsing and searching the UMLS data through the Knowledge Source Server. The three different types of client interfaces allow for a great deal of flexibility in using the system. The system can be used and explored interactively through the Web client; options can be flexibly combined for batch processing in the command-line interface; and application programs can easily extract a range of information through the API.

Large Scale Vocabulary Test

The National Library of Medicine (NLM) and the Agency for Health Care Policy and Research (AHCPR) are co-sponsoring an initiative to determine how well terms from a combination of existing biomedical thesauri reflect the terminology that is needed for developing a standard vocabulary for use in health information systems [14]. The experiment will test the coverage of some thirty thesauri that are currently represented in the UMLS Metathesaurus, together with a few that will be added for the purposes of the test. A specialized interface to the UMLS Knowledge Source Server is being developed that will be used by interested individuals who will participate by submitting their terminology to the system. Flexible access to the UMLS data will be provided through the existing indexes and through additional lexical routines. When

terms are found in the Metathesaurus, relevant data, such as definitions, semantic types, and context hierarchies will be shown to the user, so that a determination of the correctness of the match can be made.

CONCLUSION

The UMLS Knowledge Source Server is a sophisticated Internet-based tool for accessing UMLS data. It can be used by individuals who are interested in searching, browsing, or navigating through the extensive data provided by all four Knowledge Sources. Through its command-line interface, it can provide researchers with a batch processing capability for extracting and further manipulating UMLS data in their experiments. Application programs can access the UMLS data through a simple API, freeing developers from writing programs to extract the data as they appear on the CD-ROM. The benefit of this approach is that any changes to the actual UMLS data files are transparent to the application developer who, therefore, does not need to rewrite programs to accommodate the changes. A specialized interface to the UMLS Knowledge Source Server is currently under development for use in the planned NLM/AHCPR sponsored large scale vocabulary test, which should result in a better understanding of the clinical vocabulary that will be needed for a wide-range of health information systems.

REFERENCES

1. Rosenberg KM; Coultas DB. Acceptability of Unified Medical Language System terms as substitute for natural language general medicine clinic diagnoses. In Ozbolt JG (ed.), Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994:193-7.

2. Campbell JR; Payne TH. 1994. A comparison of four schemes for codification of problem lists. In Ozbolt JG (ed.), Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994:201-5.

3. Lamiell JM; Wojcik ZM; Isaacks J. Computer auditing of surgical operative reports written in English. In Safran C (ed.), Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care, 1993:269-73.

4. Burgun A; Botti G; Lukacs B; Mayeux D; Seka LP; Delamarre D; Bremond M; Kohler F; Fieschi M; Le Beux P. A system that facilitates the orientation within procedure nomenclatures through a semantic approach. Medical Informatics, 1994 Oct-Dec, 19(4):297-310.

5. Lowe HJ; Buchanan BG; Cooper GF; Vries JK. Building a medical multimedia database system to integrate clinical information: an application of high-performance computing and communications technology. Bulletin of the Medical Library Association, 1995; Jan, 83(1):57-64.

6. Cimino JJ. Use of the Unified Medical Language System in Patient Care at the Columbia-Presbyterian Medical Center. Meth Inform Med 1995; 34:158-64.

7. Robert JJ; Joubert M; Nal L; Fieschi M. A computational model of information retrieval with UMLS. In Ozbolt JG (ed.), Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994:167-71.

8. Levesque Y; LeBlanc AR; Maksud M. MD Concept: a model for integrating medical knowledge. In Ozbolt JG (ed.), Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994:252-6.

9. Peng P; Aguirre A; Johnson SB; Cimino JJ. Generating MEDLINE search strategies using a librarian knowledge-based system. In Safran C (ed.), Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care, 1993:596-600.

10. Hersh WR; Hickam DH; Haynes RB; McKibbon KA. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. Journal of the American Medical Informatics Association, 1994; Jan-Feb, 1(1):51-60.

11. Aronson AR, Rindflesch TC, Browne AC. Exploiting a large thesaurus for information retrieval. RIAO 94 Conference Proceedings, 1994; 197-216.

12. Ketchell DS, Freedman MM, Jordan WE, Lightfoot EM, Heyano S, Libbey, PA. Willow: A uniform search interface. Journal of the American Medical Informatics Association, 1996; 3(1):27-37.

13. Humphreys BL, Hole WT, McCray AT, Fitzmaurice MJ. Planned NLM/AHCPR large-scale vocabulary test: Using UMLS technology to determine the extent to which controlled vocabularies cover terminology needed for health care and public health. JAMIA, 1996; 3:281-287.